DOCUMENT RESUME

ED 262 041 TM 850 489

AUTHOR Diamond, Esther E.
TITLE Content, Context, and Construct Considerations in Sex Bias in Testing.
PUB DATE Apr 85
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; College Entrance Examinations; Content Analysis; Elementary Secondary Education; *Experience; Higher Education; Literature Reviews; Occupational Tests; Research Needs; *Sex Bias; Sex Differences; *Sex Fairness; Sex Role; *Test Bias; *Test Construction; Test Items; *Test Validity

ABSTRACT

Studies of sex differences in test performance have identified the content of test items, the context in which they are set, and the constructs purportedly underlying the item as important variables influencing performance. Test content often reflects experiences that traditional social roles have discouraged women from exploring. Often, the item pool in an occupational area dominated by men tends to be based on male experiences or to use male referents almost exclusively. The same is true for roles in which females predominate, but the domain is narrower than that for men. Mathematics and science items presented in a context more familiar to men than to women have been found to be more difficult for women. In examining item content, it is essential to identify content that is irrelevant to the intended construct but that might elicit different responses to items from different groups. Further research is needed on the differences between judgmental and statistical methods of item bias detection and on the types of bias detected by different methods. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262041

Content, Context, and Construct Considerations in Sex Bias in Testing

Esther E. Diamond

Educational and Psychological Consultant

Evanston, Illinois

BEST COPY AVAILABLE

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1985.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E. E. Diamond

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 489

Abstract

The content of test items, the context in which they are set, and the constructs purportedly underlying the item have been identified as important variables associated with male-female differences in test performance. Test content often reflects experiences that traditional social roles have closed to women or discouraged women from exploring. Often, the item pool in an area traditionally considered one in which males are dominant tends to be based on male experiences or to use male referents almost exclusively. (The same is true for roles in which women predominate, but the domain is much narrower than that for men). For example, mathematics and science items presented in a context more familiar to men than to women, for whatever reason, have been found to be more difficult for the latter. In one study, vocabulary items biased in favor of men were scientific or practical, while those biased in favor of women were person-oriented. In examining item content, as Cole (1981) pointed out, it is essential to identify content that is irrelevant to the intended construct but that might elicit different responses to items from different groups. This paper examines relevant studies on sex differences in test performance and suggests needed research.

BEST COPY

BEST COPY AVAILABLE

Content, Context, and Construct Considerations in Sex Bias in Testing

Esther E. Diamond

Educational and Psychological Consultant

Evanston, Illinois

Item content, context, and contruct are overlapping considerations in sex bias in testing. The content of a test item may be presented without a defined context--as, for example, a straighforward computation item or a vocabulary item that is not presented in a phrase or sentence. Most test items, however, are embedded in a particular context, and it is these two aspects of an item--content and context-- that are most closely related to facial bias.

Facial bias occurs when certain aspects of item content and the context in which the item appears are judged to have a potentially negative effect on the performance of that group. Such aspects might include exclusive use of male referents, or depiction of members of a particular group in stereotypical roles or behaving in stereotypical ways, or use of situations or item formats unfamiliar to members of that group. Whether such items should be retained even if they do not affect test performance has been answered by Cole (1981), Messick (1980), and Diamond and Tittle (1985), among others: Items that are judged to portray members of a group in stereotypical fashion or that are in any way demeaning to a group should not be included in a test.

Content/Context Effects on Performance

The research on how item content affects performance shows somewhat mixed results. Not all studies showed differences. Plake, Hoover, and Loyd (1978), for example, investigating differential item performance by sex on the ITBS, found an item-by-sex interaction that was statistically but not practically significant, with a minimal percentage of the total variance accounted for by the interaction components. Scheuneman (1984) maintained that item bias studies have failed to substantiate the argument that bias in items is the result of test content differentially familiar to persons from different groups.

On the other hand, a number of studies do show that content and context affect male and female test performance differentially. For example, Donlon, Hicks, and Wallmark (1980) studied sex differences on the Graduate Record Examination (GRE). They used basic correlations between item difficulties for males and females as the major test of equivalence of test material for the two groups. They found a high order of consistency in item difficulty--that is, items were about as difficult or as easy relative to each other within the male groups as within the female groups, with slope parameters approximating unity in each case. Most marked deviations from unity were for Sentence Completion and Data Interpretation (males did better on more difficult items in these tests) and Reading (females did better on more difficult items). The investigators also found that the content of male-biased vocabulary items was more scientific or practical, while the content of female-biased items was more

BEST COPY AVAILABLE

person-oriented. They concluded that item types per se (an aspect of context) are not linked to relative difficulty for the sexes, but that item content is a determining factor.

In Math, the investigators hypothesized, data interpretation items were relatively more difficult for females because males receive more practice in graphic representation (context again). They concluded that it is possible to influence the magnitude of observed differences between the sexes by controlling the selection of item material, although they conceded the importance of measuring skills such as the handling of complex graphs as a bona fide outcome of education. They urged that observed differences be examined from the standpoint of educational background.

Moulton, Robinson, and Elias (1978) investigated whether the use of male terms in their "gender-neutral" sense would induce people to think of males even in contexts that were explicitly gender-neutral. The investigators found a significant tendency for more male exemplars to be selected as the fictional character of a story in response to a male term (his) used as a generic, or gender-neutral pronoun than in response to a genuinely neutral term (e.g., he or she, his or hers, their), even for explicitly gender-neutral contexts.

Jensen (1980) cited McNemar to the effect that large differences in percent passing an item can usually be accounted for in terms of experiences or training typically associated more with one sex than with the other and that, in the Stanford-Binet:

> no one type of item consistently favors either sex. The sex differences in item difficulty are apparently more a function of specific item content than of the types of basic abilities

BEST COPY AVAILABLE

called for by the item (p. 623).

We would interpret "types of basic abilities" as constructs the test is purportedly measuring.

Studies by Medley and Quirk (1974) and Donlon, Ekstrom, and Lockheed (1979), also indicated that test content differentially affects the performance of males and females on achievement tests.

While, as indicated earlier, content and context in test items are closely intertwined, a number of studies have identified item context as mainly responsible for male/female differences in test performance. In fact, in mathematics and science tests context may be one of the most important variables affecting performance. Milton (1958), for example, investigated sex-role identification and problem-solving achievement of high school boys and girls. When scores were adjusted on the basis of a sex-role identification measure, problem-solving differences between males and females were nonsignificant. In a second study, problems were presented in a content/context framework oriented to the female role, and differences in performance were reduced further.

Similarly, McCarthy (1975) found more systematic differences in performance on test items judged to be more familiar to females or to males, or equally familiar to each in the context or setting for problem-solving items. She used mathematics content and a sample of high school students in grades 10 to 12. She constructed an item pool, varying the item context and holding the mathematical processes constant. Her results were more in line with the general trend in the research literature toward no sex differences in math performance in the early grades and some findings of sex differences in the later,

BEST COPY AVAILABLE

high school grades.

Experience and Context

Test content and context often reflect experiences that traditional social roles have closed to women or have discouraged them from exploring. One might ask, for example, whether differences in male and female responses to particular items are directly related to the construct that is purportedly being measured or to the context in which the purported construct is presented. Many women, for example, may not have had any experience with such tasks as blueprint reading or understanding how a carburetor works. Seeing the spatial relations between parts of a pattern, however, would seem to require skills similar to those required for blueprint reading, and understanding the way a sewing machine works may well involve the kinds of concepts involved in understanding the workings of a carburetor.

Anastasi (1980) pointed to an increasing body of data indicating that it is not only the level of performance in different abilities, but also the way in which performance is organized into distinct traits, that is influenced by experiential background. It seems obvious that the experiences that males and females are exposed to throughout their lives--including the courses they take in school, the activities they engage in, and the attitudes of parents, teachers, and society in general--affect the ways in which they respond to different contexts. This can be seen in the studies cited above, and it is dramatically illustrated in a study by Carlson (described in Cole, 1978). Carlson examined items categorized by content on the science achievement test of the International

BEST COPY AVAILABLE

Association for the Evaluation of Educational Achievement (IEA). Among the sex differences in item responses was an "understanding" item involving how to put batteries into a flashlight; boys did better. Girls, on the other hand, did better on another "understanding" item--how to get the lid of a jar off by placing the jar under hot water.

If the item context reflects traditionally male experiences to which most women have had little or no exposure, it is not surprising that males do better. The reverse would be expected if the context reflected traditionally female experiences. We see this point illustrated, too, in interest inventories and vocational aptitude tests, where items contributing to scores on scales for traditionally male occupations or areas of interest are often presented in contexts reflecting experiences to which on sex or the other has had little exposure.

Problems in Search of Solutions

The consciousness of those who use tests, those who take them, and--perhaps most importantly--those who develop and publish them has been raised considerably in the past two decades. Most major test publishers have developed and are using bias-detection guidelines and/or statistical procedures for detecting bias (see, for example, Berk, 1982). A followup survey of an earlier one (Diamond, 1980) on sex bias in achievement testing, just completed by the Commission on Bias in Testing of the Association for Measurement and Evaluation in Counseling and Development, indicates that most new tests and revisions of older tests are undergoing careful scrutiny for bias. Regulations for Title IX of the Education Amendments of 1972 have, at

BEST COPY AVAILABLE

least until now, provided an effective goad for such scrutiny.

Recommendations for research, as well as for test use, regarding sex bias in tests are given in Diamond and Tittle (1985).There are, however, problems for which there does not seem to be a ready solution, and much research remains to be carried out. Primarily, we need to be sure that when item content or context is changed, the item is still measuring the intended construct. Do blueprint reading and pattern manipulation measure the same skill or ability? Is understanding how a sewing machine works measuring the same construct as is measured by understanding how a carburetor works?

Cole (1981) has observed the need for judgments different from judgments of facial bias. Facial bias, she pointed out, since it may not have any impact on subjects' responses to test items, may be unrelated to construct validity bias detected by statistical item bias procedures. The need, then, is for:

> ...judgments involving characteristics of item content that are irrelevant to the intended construct, but that might elicit different responses to items from different groups of subjects. Such construct-validity-based item judgments seem to require knowledge of the construct being measured by the test as well as knowledge of different groups being studied. It is likely that such judgments involve much more difficult and subtle types of judgment than facial bias judgments (p. 1073).

Cole (1981) also stressed the necessity for examining whether the internal test structure of items or subscores is similar for different groups:

BEST COPY AVAILABLE

> If scores in different groups yield different interrelationships, then the test is behaving in a way that would not be expected from the construct supposedly being measured, and hence the possibility of invalidity (bias) with respect to some group would be raised (p. 1071).

Cole suggested use of factor analysis to describe interrelationships among several variables in terms of a smaller number of dimensions; if dimensions are different for different groups, this may be seen as a type of bias.

Another problem is that research comparing the data from judgmental and statistical methods of item bias detection have found almost irreconcilable differences between the methods (Berk, 1982). Berk suggested that a more perceptive interpretation might be that the two methods are tapping different kinds of bias. He recommended integration of the information from both types of analysis in order to gain as much insight as possible into the various identifiable forms of bias and to satisfactorily address all claims of bias.

Scheuneman (1984) pointed out that judgmental methods have had far more impact than statistical methods--that the latter have typically not had results that could be generalized into guidelines for test developers. Like Berk, she suggested that both methods may be detecting fundamentally different processes. What these processes are is a major area for research.

The concerns, methodology, and research discussed here must be part of a dynamic process in the effort to detect and minimize item bias. The process must be carried on with the involvement and cooperation of all possible stakeholders--from the test developers on

BEST COPY AVAILABLE

down to those who are examined by the tests and whose education and careers are influenced by the scores they attain.

BEST COPY AVAILABLE

References

Anastasi, A. (1980). Sex differences: Historical perspectives and theoretical implications. Psychological Documents, Ms. 1999. Washington, D.C.: Americocan Psychological Association.

Berk, R.A. (Ed.) (1982). Handbook of methods for detecting sex bias. Baltimore, Md.: Johns Hopkins University Press.

Cole, N.S. (1978, March). Approaches to examining bias in achievement test items. Paper presented at the annual meeting of the American Personnel and Guidance Association, Washington, D.C.

Cole, N.S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Diamond, E.E. (1980). The AMEG Commission report on sex bias in achievement testing. Measurement and Evaluation in Guidance, 13, 135-147.

Diamond, E.E., & Tittle, C.K. (1985). Sex equity in testing. In S. Klein (Ed.), Handbook for achieving sex equity through education (pp. 167-188). Baltimore: Johns Hopkins University Press.

Donlon, T.F., Ekstrom, R.B., & Lockheed, M.E. (1979). The consequences of sex bias in the content of major achievement test batteries. Measurement and Evaluation in Guidance, 11, 202-216.

Donlon, T.F., Hicks, M.M., & Wallmark, M.M. (1980). Sex differences in item responses on the Graduate Record Examination. Applied Psychological Measurement, 4, 9-20.

Jensen, A. (1980). Bias in mental testing. New York: The Free Press.

McCarthy, K. (1975). Sex bias in tests of mathematical aptitude. Unpublished doctoral dissertation, City University of New York.

BEST COPY AVAILABLE

Medley, D.M., & Quirk, T.J. (1974). The application of a factorial design to the study of cultural bias in the General Culture items in the National Teacher Examination. Journal of Educational Measurement, 11, 235-245.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Milton, G.A. (1958). Five studies of the relation between sex-role identification and achievement in problem solving. Technical Report 3. New Haven, Conn.: Yale University.

Moulton, J.,Robinson, M., & Elias, C. (1978). Psychology in action: Sex bias in language use--"Neutral" pronouns that aren't. American Psychologist, 33, 1032-1036.

Plake, B.S., Hoover, H.D., & Loyd, B.H. (1978, March). An investigation of differential item performance on the Iowa Tests of Basic Skills. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Scheuneman, J.D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. Educational Psychologist, 19, 219-225.

BEST COPY